Rich Representations of Visual Content for Screen Reader Users

Meredith Ringel Morris[1], Jazette Johnson[1,2], Cynthia L. Bennett[1,3], Edward Cutrell[1]
[1]Microsoft Research, Redmond, WA, USA
[2]Vanderbilt University, [3]University of Washington
{merrie, cutrell}@microsoft.com, jazette.johnson@vanderbilt.edu, bennec3@uw.edu

ABSTRACT

Alt text (short for "alternative text") is descriptive text associated with an image in HTML and other document formats. Screen reader technologies speak the alt text aloud to people who are visually impaired. Introduced with HTML 2.0 in 1995, the *alt* attribute has not evolved despite significant changes in technology over the past two decades. In light of the expanding volume, purpose, and importance of digital imagery, we reflect on how alt text could be supplemented to offer a richer experience of visual content to screen reader users. Our contributions include articulating the design space of representations of visual content for screen reader users, prototypes illustrating several points within this design space, and evaluations of several of these new image representations with people who are blind. We close by discussing the implications of our taxonomy, prototypes, and user study findings.

Author Keywords

Alt text; alternative text; captions; screen readers; visual impairment; blindness; accessibility.

ACM Classification Keywords

K.4.2. Assistive technologies for persons with disabilities.

INTRODUCTION

Digital imagery pervades modern life. More than a billion images per day are produced and uploaded to social media sites such as Facebook, Instagram, Flickr, Snapchat, and WhatsApp [9]. Beyond social media, we also encounter digital images within websites, apps, digital documents, and electronic books. Engaging with digital imagery is part of the fabric of participation in contemporary society, including education, the professions, e-commerce, civic participation, entertainment, and social interactions.

Most digital images remain inaccessible to people who are blind. As of 2014, the World Health Organization estimates that 39 million people are blind, and 246 million have low vision [36]. People who are blind use screen reader software to operate their computers and mobile devices. Most major operating systems come with built-in screen readers that can be enabled in the accessibility settings (e.g., Apple's VoiceOver, Google's ChromeVox and TalkBack, Microsoft's Narrator), and many people also choose to install third-party screen readers such as JAWS or NVDA. Screen readers render on-screen text as audio, and the user can navigate among different parts of the interface using shortcut keys (on a desktop or laptop computer) or gestures such as taps or swipes (on a tablet or smartphone).

Screen readers cannot render an image as audio unless the content author has specified *alternative text* (also called *alt text*) for that image [35]. If no alt text is present, the screen reader may simply announce "image" or skip the image entirely; if an alt text is present, it will be read aloud. Most digital platforms offer a way to provide alt text, whether as a property that a programmer can specify when writing software for various operating system platforms, an HTML attribute when authoring a web page, or an attribute that can be added via a context menu when authoring various document types such as word processor documents and slide decks. In each case, the alt text consists of a descriptive caption in the form of a short phrase or sentence which, if present, is read aloud by the screen reader when it encounters that image.

Alt text arose with the HTML 2.0 standard in 1995 [1, 3], wherein the "img" tag used to place images within HTML documents allowed for an "alt" attribute that specified text that could be rendered in the case that the image could not be (see the Appendix for an example). While this has come to be primarily used by screen readers, this property was originally conceived of for use in cases where the user had a text-based Web browser such as Lynx [lynx.browser.org/] or had a very slow internet connection, in which case the alt text could be rendered temporarily until the complete image managed to download [16].

The capabilities of computers and the volume, importance, and purpose of digital imagery have evolved substantially since 1995. The experience of consuming visual content via a screen reader, however, has remained frozen in time. In this paper, we consider how modern computing capabilities such as interactivity, high-fidelity audio capabilities, touch interaction, and real-time crowdsourcing [4] and friendsourcing [7] can provide a new experience of visual content for screen reader users.

In this paper, we use the term "alt text" to refer to the information used to convey visual content to a screen reader

CHI 2018, April 21–26, 2018, Montreal, QC, Canada © 2018
Copyright is held by the owner/author(s).
Publication rights licensed to ACM.
ACM 978-1-4503-5620-6/18/04…$15.00
https://doi.org/10.1145/3173574.3173633

user, since alt text is the most common status quo today for representing such content (though other markup, such as ARIA [37], is becoming increasingly important). However, supplementing or modifying the alt text element in HTML or XML is not necessarily the optimal way to represent or deliver the experiences we propose. Exploring the pros and cons of different formats and standards for non-visual image representations (and considering the impact that any such standard may have on content authorship and accessibility compliance) is an important avenue for future research that is beyond the scope of what we address in this paper.

After discussing related work, we introduce a taxonomy of properties that create a design space of possibilities for non-visual image representation. We then describe a series of prototypes we developed that instantiate interactions illustrating different combinations of properties within this design space. We evaluated several of these novel interactions with fourteen blind screen reader users, and report quantitative and qualitative findings from these user tests. Finally, we reflect on the implications of this work for improving image accessibility, and identify key considerations going forward.

The contributions of this research include: (1) a design space for representations of visual content for screen reader users; (2) prototypes of novel interactions that supplement the default alt text experience; and (3) feedback from screen reader users on these novel experiences, resulting in implications for design.

RELATED WORK

Prior research and standards on labeling images for accessibility purposes focus on two main areas: (1) guidelines and end-user preferences concerning captions, and (2) methods for generating captions when content authors fail to do so themselves.

Captioning Guidelines and Preferences

The World Wide Web Consortium (W3C) provides guidelines for the composition of image captions for screen reader users in the WCAG [39]. These guidelines state that all non-text should have a description, but some debate about which images should have descriptions and how the descriptions should be composed has ensued. Some guidelines such as those by Slatin and Rush [28] state that descriptions should make sense in and out of context of surrounding content. This means that an author should not rely on inline text to help someone understand the photo; rather, it should be understandable even if the photo description is the only thing someone reads. However, several compliance tutorials including those by the WCAG [40] and Section 508 [12] advise that if inline text describes the image, further image description is redundant.

Petrie et al. [23] interviewed five blind people to learn their preferences for image descriptions. They reflected existing guidelines in that most did not want descriptions of decorative images, spacers, or logos. Participants wanted a variety of components to be described, emphasizing context as the most important determiner of what should be prioritized. In general, they wanted to know the purpose of the image, what objects and people are present, any activities in progress, the location, colors, and emotion. Participants agreed on four types of image that they would want described: products for sale, art, images on functional components of a web page like buttons or links, and descriptions of graphs and charts. When asked how long a description should be, participants agreed that a few words were not usually enough and that they would read a long description if one was available. They did, however, emphasize that the description should present the most important information first.

Morash et al. found that online workers did a poor job when asked to create alt text for STEM images (e.g., charts and graphs for science textbooks) according to accessibility guidelines; however, they found that workers produced higher quality output when they used a template that required them to fill in specific details that could then be combined to create a caption [20]. Salisbury et al. [27] developed a set of structured questions about social media images that can guide humans or AI systems in creating captions that contain the types of details desired by people who are blind.

There is debate over the length of captions, with many guides advocating brevity (e.g., [28, 35]) though some research suggests users prefer detail [23, 27]. HTML 4 introduced the *longdesc* attribute on the *img* tag, which could supplement the *alt* attribute; whereas the alt text provides a brief description, for complex images the longdesc attribute could point toward a separate URL that would contain more lengthy details. However, the longdesc attribute was rarely used; a 2007 analysis found that of one billion online images, fewer than 0.13% included the longdesc attribute at all, and of those images using the attribute, 96% were misusing it (e.g., leaving it blank, pointing to an invalid URL, pointing to the image itself, etc.) [24]. The longdesc attribute remains controversial: it was deprecated and removed from the HTML 5 standard [30], but was later reinstated [38].

These guidelines for best practices in alt text composition and studies about the level and types of detail preferred by screen reader users informed our prototypes. We aim to offer a supplement to alt text that allows access to richer detail than the status quo, but that also offers the end user control over the time spent interacting with an image. Our present work is focused on representations of the visual, rather than functional, aspects of online imagery (e.g., our taxonomy and prototypes did not explore imagery used for organizational or navigational purposes).

When Alt Text is Missing

Many web pages, apps, documents, and social sites contain low-quality alt text (e.g., a filename like “img123.jpg” or the word “image” as the alt attribute) or have no content at all for the alt attribute [5, 13, 17, 18, 21, 22, 25, 29, 34]. Some platforms and apps still fail to include any method at all for specifying alt text; for example, the popular social media platform Twitter, despite having images in about one-third of its posts by 2015 [21], did not even add the capability for users to specify an alt attribute until 2016 [32]. New standards such as ARIA hold promise for supporting labeling of an expanded set of web elements, including dynamic content [37].

Researchers have explored the applicability of human computation techniques to produce near-real-time image labels via either crowdsourcing [4, 33, 42], friendsourcing [6], or social microvolunteering [7]. However, these incur monetary, privacy, and accuracy costs (in the case of paid crowdsourcing), social costs (in the case of friendsourcing), or scalability concerns (in the case of social microvolunteering). Combinations of human-in-the-loop techniques with automated techniques like OCR can help backfill missing captions in some cases (e.g., the WebInSight system [5]).

Recent advances in computer vision technology (e.g., [10, 31]) have made completely automated captioning feasible in some cases (e.g., high-quality images of certain types of objects); in 2016 Facebook introduced an Automatic Alternative Text feature that applies tags to images in users' feeds identifying objects within them (e.g., "this image may contain a dog, trees, the sun") [41]. However, computer-generated captions are often inaccurate, and users who are blind tend to place more trust in such systems than is warranted by the state of the art [19]. Further, automated captions do not yet include the amount and types of detail desired by many screen reader users [27].

In this work, our focus is not on the challenge of supplying missing alt text, but rather on considering novel representations of visual content that can offer screen reader users a richer understanding of digital imagery, assuming caption content is available. Our proposed interactions are agnostic to authorship; while we primarily envision the requisite details being supplied by content authors, automated or human-in-the-loop techniques could also be substituted.

DESIGN SPACE

We propose a taxonomy of properties relevant to representing visual content non-visually (i.e., to screen reader users). Articulating this taxonomy reveals an unexplored design space of property combinations. This design space can be a useful tool for researchers interested in expanding the possibilities for presenting image captions to screen reader users. Our taxonomy comprises five categories: *interactivity*, *stability*, *representation*, *structure*, and *personalization*.

Interactivity

The interactivity category indicates whether an alt text is *passive* or *active*. Standard alt text is passive; the screen reader simply reads them aloud to the user. However, there is no technical impediment to developing sysetms in which the user's activity (e.g., spoken commands, key or button presses, touch or gesture interactions, etc.) might determine the alt text they receive.

Stability

The stability category indicates whether an alt text is *static* or *evolving*. Standard alt text is static; the content author (or AI system) produces a single caption for the image. However, there are many reasons why a caption might evolve: one may wish to add additional content requested by an end-user, correct mistakes made by an automated system, or include user feedback on caption quality.

Representation

The representation category indicates the format(s) in which caption information is presented. Standard alt text consists only of *text* (which is rendered aurally to screen reader users with text-to-speech functions). However, other media formats or modalities could also be used to represent a rich and evocative understanding of the image to the end user, including *sound effects*, *music*, *sonification* (in which visual properties such as color are mapped to audio properties, e.g. [26]), *vibration*, *haptics*, etc. Some types of feedback may require specialized hardware which may not yet be mainstream (e.g., tablets capable of producing spatialized haptic sensations), but considering both current and not-yet-possible representations lets us design systems with the flexibility to suit future output technologies as they come into being.

Structure

The structure representation category indicates whether an alt text is *structured* (i.e., contains semantic metadata) or *unstructured*. Standard alt text is unstructured, consisting of one or more words, phrases, or sentences. However, structuring alt text according to an ontology would allow users to query or filter on certain types of detail based on their interests. Different types of images may warrant different ontologies (e.g., some categories that might apply to graphs and charts may not apply to selfie photos, etc.); structure might also confer authorship benefits by providing guidance about what categories of information should be included in an alt text (e.g., [20]).

Personalization

The personalization category designates whether the alt text is *generic* or *personalized* to a specific end-user or group. Standard alt text is generic; the same alt text is read to any user who encounters the image. However, it may be desirable to personalize aspects of alt text delivery. For example, caption presentation could be personalized for social media images based on a user's relationship with a person depicted in a photo (an alt text might describe a person as "your sister" to one user and "Jane Smith" to another, depending on relationship information). Caption delivery could also be personalized based on a preference profile the user establishes with a screen reader, such as identifying the types of details that most interest them or their preferred language. A user's interaction history could also personalize caption

Table 1. This table shows how our novel interactions, as well as standard alt text, fit into our proposed design space.

"Alt Text" Style	Interactivity	Stability	Representation	Structure	Personalization
standard	passive	static	text	unstructured	generic
progressive detail	active	static	text	structured	generic
multimedia	passive	static	text + music/sound effects	unstructured	generic
spatial	active	static	text + kinesthetic feedback	unstructured	generic
categories	active	static	text	structured	personalized
question & answer	active	evolving	text	structured	generic or personalized
hyperlink	active	static	text + kinesthetic feedback	structured	personalized

playback; for example, perhaps the first time a specific user encounters an image they would hear a longer and more detailed caption than on subsequent encounters with that same image.

PROTOTYPES

Drawing inspiration from our design space, we prototyped six novel interactions that illustrate different properties of this taxonomy: Progressive Detail, Multimedia, Spatial, Categories, Question & Answer, and Hyperlink. We implemented our prototype as an Android application, using a seven-inch Android Tablet running the TalkBack screen reader. We used buttons to trigger our alt text interactions (except in the "spatial" and "hyperlink" techniques, where we rely on direct-touch); however, these techniques are applicable to other form-factors (e.g., laptops, phones) and could be triggered by other means (e.g., voice input or keyboard shortcuts instead of buttons, mouse clicks or arrow keys instead of direct touch). The accompanying Video Figure demonstrates each of the six interaction styles in our prototype, which we describe below. Table 1 shows how our prototype techniques fit into the design space taxonomy articulated in the previous section. The Appendix shows the XML syntax we used when creating our prototypes.

Progressive Detail

The *progressive detail* alt text is designed to give the user control over the level of detail received based on their interest in the image. The content author can specify multiple alt texts for the same image, and indicates their logical ordering (first, second, third, etc.). The first of the progressive detail alt texts is meant to be equivalent in detail to a standard alt text, while subsequent ones may reveal more information. Our implementation can support a variable number of detail levels; for our evaluation with end-users, we used three levels of detail on all images for consistency during user testing. We considered different interactions for accessing these details: in one design, we created buttons the user could use to increase or decrease the current level of detail before playing the alt text; for our user study, we used a variation in which each level of detail could be accessed specifically with its own button (which works well for a three-level alt text, but may not scale well depending on the number of detail levels). Another design decision is whether levels of detail should be independent of one another, or whether higher levels should recap the contents of lower levels before adding additional content; this latter design would allow a user to select a single level of detail (either on the basis of a specific image or as a general setting on their screen reader) and hear a single alt text description at the desired level, whereas the more independent detail levels support more *ad hoc*, progressive exploration while minimizing redundancy. We used the independent detail design for our user testing.

Multimedia

The *multimedia* interaction style introduces the concept of "alt audio" to supplement (or possibly substitute for) alt text. The *altAudio* attribute of the *img* tag specifies an audio file, such as music or sound effects. We envision this technique being used to create a rich sense of presence and aesthetic around an image. For example, perhaps a traditional wedding march song might play in the background as the alt text description is read for a photo of a bride and groom walking down the aisle, or a photograph of Fourth of July fireworks might include sound effects of the explosions that accompanied the visuals. Such sounds might be curated by the author from personal libraries or publicly available sources, suggested (or even composed) by future AI algorithms based on visual or metadata properties of the image, or even captured in-situ by novel photography tools (prototype photo applications for blind users like VizSnap [1] support capture of ambient audio and Accessible Photo Album [14] supported audio replay; audiophotography [11] could become part of mainstream photography). When prototyping the multimedia interaction, we explored several possibilities for how a screen reader user might consume this alt text – for example, we created implementations in which the alt audio and alt text were overlaid and rendered simultaneously (we used volume mixing, but one might also use 3D audio technology to improve the user's ability to discern both tracks), as well as implementations in which the user could request and listen to the verbal and nonverbal information separately. For the user study, we used the simultaneous-presentation implementation, since this was most distinct from the standard experience.

Spatial

The *spatial* interaction style is designed to allow direct interaction with an image in a manner that can help the user build a mental model of the relative locations of components of the image. Within the *img* tag, any number of polygonal regions can be defined (our prototype uses rectangles), and a separate *alt* attribute associated with each. The associated audio can be played when the user touches the specified region of the image when rendered on a touch-screen device;

mouse-clicks could trigger a similar interaction on a traditional computer. If the image is not rendered full-screen, the TalkBack screen reader plays a "click" sound to alert the user when their finger drifts beyond the picture's boundaries.

Categories

The *categories* technique creates a highly-structured set of metadata that can be queried as a supplement to a traditional, free-form alt text. This technique takes inspiration from findings that structured forms can help workers create better alt text for STEM diagrams [20]. Our prototype enables six categories of information that Salisbury et al. [27] found were important details to include in captions of social media images for people who are blind; categories that are not applicable to the current image receive a default value such as "none," "unknown," or "not applicable." *Image type* indicates the class of image (e.g., a snapshot, a formal portrait, a cartoon, a painting, a logo, a graph, a meme). *Setting* describes the location of the image (e.g., at the beach, in a kitchen, on the White House lawn). *Number of people* indicates how many people are shown in the image; for images of large crowds, this might be an estimate (e.g., hundreds). *Action* describes the primary event represented in the image (e.g., dancing, protesting, smiling). *Emotion* conveys the aesthetic feeling produced by the image (e.g., gloomy, celebratory, nostalgic). *Humor* indicates whether the image is intended to be funny. This set of initial categories is intended to illustrate the concept of a highly-structured alt text; we anticipate that the specific categories included might vary depending on the type of image (e.g., a diagram versus a photo) and the context in which it appears (e.g., a shopping site versus a social media post). A user profile can personalize delivery by specifying which categories interest a user, and these categories of interest are always read aloud if present, while others are delivered only when the user queries them; if no user profile is provided, then categories are only described when queried.

Question & Answer

Our *question and answer* alt text allows for content that grows over time. If the listener has a question about the image or wants to know details not present in the base alt text, they can press a button that allows them to enter a question. Our prototype supports two methods of entering questions, either by typing or by using voice input that is translated into text by Android's automatic speech recognition functionality. Questions, and any answers eventually received, are appended to the alt text's XML such that when the user hears the base alt text they are also told how many questions are present; the user can choose to play each of the questions, and, if an answer has been supplied, it is read aloud after its question (otherwise the system indicates the question has not yet been answered). In this way, one user may benefit from hearing the details requested by a different user; alternatively, this alt text style can be personalized by only including questions from the current user (and the associated answers). In our prototype, we edited answers directly in the XML; in a deployed system, we imagine that notifications of questions could be sent to the content author who could perform such edits. Alternatively, questions could be sent to and answered by the crowd [4] or friends [6, 7].

Hyperlink

The *hyperlink* interaction is meant to allow users to interactively query detailed visual descriptions or other details related to well-known people, places, or things that may be present in an image. When reading the alt text, we play a sound effect when an object with available hyperlink detail is announced, and the user can press a button if they wish to hear that detail. For example, a photo of the Washington Monument might have an alt text that reads "A photo of the Washington Monument," which is reasonably informative, but someone who is blind may know conceptually that the Washington Monument is an important piece of architecture but not have a sense of what it looks like. If they want to hear a visual description of that monument, they could use the hyperlink interaction to hear a visual description, such as "The Washington Monument is a white marble obelisk standing 555 feet tall." For our prototype, we curated a database of descriptions for objects appearing in our sample set of photos; we envision the hyperlink interaction could be created at scale by using an entity extractor on alt text descriptions, checking which entities had Wikipedia pages, and reading descriptions from those pages; crowd workers or volunteers might then refine these descriptions as needed. Hyperlink descriptions could also be personalized to include a database of entities specific to a user, such as physical descriptions of contacts from their own social network or local points of interest; a database of such personalized descriptions could be generated via friendsourcing or crowdsourcing, or even using computer vision as that technology's accuracy evolves. In our prototype, we combined the hyperlink interaction with the spatial technique, so that as the user's finger drags over an object for which a hyperlink is available, they can double-tap that region of the image to hear an additional description.

EVALUATION

We conducted a user study to better understand how screen reader users would value the expansions to alt text conceptualized in our taxonomy and prototype. We had to make a trade-off to limit the total number of interaction techniques that we evaluated for the available time. To avoid fatiguing participants, we limited the session length to one and one-half hours, but we wanted to evaluate each technique in sufficient depth to provide participants a robust experience of each in a range of potential contexts of use. Therefore, we selected three of our prototype interactions that represented different characteristics from the design space: multimedia, spatial, and progressive detail. The study aims to answer the following research questions about these three interactions:

RQ1: Which interaction styles do screen reader users prefer, and do these preferences vary for different use scenarios (e.g., reading the news, consuming social media, shopping online, studying from a digital textbook)?

RQ2: How do these new interaction styles impact users' understanding of images?

RQ3: What do screen reader users view as the benefits and limitations of these new interaction styles? How can these interactions be improved?

Participants

We recruited fourteen legally blind adults from the Seattle metropolitan area using advertisements on social media and sent to email lists for local organizations related to visual impairment. Participants came to our lab for between one to one-and-a-half hours to complete the study, and were paid $100 for their time. Participants' ages ranged from 25 to 65 years old (mean = 43.7 years), and gender was evenly split between male and female. Most participants described themselves as either completely blind or having only light perception, but six described having very small amounts of residual vision (low enough to meet the definition of legal blindness); all of the participants relied on either a guide dog or white cane to navigate the physical environment, and all used screen reader technology for interacting with computers.

Method

We began the study sessions by asking participants a few questions about their background (age, gender, description of their level of vision, preferred screen reader software). Next, we explained that we would be showing them several prototypes embodying new formats and interactions for alt text, and that these prototypes would use an Android tablet running the TalkBack screen reader. We explained that the use of the tablet form-factor and the use of buttons to trigger the interactions were only the instantiation used in our current prototype, and that in the future other interactions (such as gestures, keyboard shortcuts, or voice commands) might be used instead of software buttons. We provided a brief tutorial of the gestures that would be needed to operate our prototype using TalkBack. Because all participants were screen reader users and were familiar with alt text interactions from their daily computing experience, we provided a brief tutorial demonstrating how a standard alt text could be played using our system before focusing the remainder of the session on the three novel interaction styles (multimedia, spatial, and progressive detail).

The three interaction techniques were presented to users in a counter-balanced order using a Latin Square design, to mitigate order effects. For each of the three interaction techniques, we first presented a tutorial explaining the technique's properties and allowing the user to practice using the technique on a sample image until they were satisfied they understood it. Then, we presented the user with four different scenarios: studying from a digital textbook, browsing social media, reading a news website, and shopping online. For each scenario, the user experienced an image using the current interaction technique, taking as much time as they needed to explore the image.

They experienced a different image paired with each scenario for each of the three techniques so that they had a new image to explore each time. Images for each scenario were selected so as to be similar in nature (e.g., images containing people from NBC News' "top images of 2016" list for the news scenario, images featuring a single gender-neutral fashion accessory for the shopping scenario, images featuring a famous American monument for the textbook scenario, and images featuring people engaging in endearing behaviors for the social media scenario; the mapping of image to technique was also counterbalanced to mitigate any effects of specific images.

After each scenario, we asked participants a series of Likert-type questions to gauge their confidence that they understood the image, their satisfaction with the level of detail provided, and their rating of the suitability of the current technique for the current scenario. After completing all four scenarios with a given technique, we asked an additional set of questions asking the user to rate their satisfaction with the technique more generally, and to describe what they liked and disliked about the technique, including how they thought the technique could be further improved. Finally, after experiencing all three techniques (multimedia, spatial, and progressive detail), participants were asked to rank these three techniques together with standard alt text in order of preference. We asked them to comment on the rationale behind their ranking, and to offer any final comments or ideas about any of the techniques, as well as any other ideas they had about ways to improve the experience of consuming alt text.

FINDINGS

Here, we report on the findings from our evaluation, organized by our three primary research questions. We used non-parametric statistical tests on participants' Likert-type question responses, due to the ordinal nature of such scales. All significance information incorporates Bonferroni corrections for multiple comparisons.

RQ1: Preferences for Interaction Styles

After experiencing each of the four scenarios (digital textbook, social media, news, and shopping) with each technique, participants were asked to rate their agreement with the statement "I would use <technique X> when consuming images in <scenario Y>" on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree). Table 2 summarizes these ratings. For the progressive detail technique, there was a significant difference in ratings of likelihood to use the technique in different scenarios, $\chi^2(3, N=14) = 16.81$, p = .001, although follow-up pairwise comparisons were not significant for any of the scenario pairs after Bonferroni corrections were applied. The spatial technique did not have significant differences in ratings of likelihood of use across scenarios. The multimedia technique did have a significant difference in ratings of likely use for different scenarios, χ^2(e, N=14) = 11.14, *p* = .01. Follow-up pairwise Wilcoxon tests indicate that users were more interested in using multimedia alt text for the news scenario than for the digital textbook scenario (*z* =-1.46, *p* = .02).

To compare the interest in using each technique more generally, we took the median rating of their desire to use each technique across all four scenarios. The median Likert rating for the progressive detail technique was 5, for the spatial technique was 5, and for the multimedia technique was 2.75. A Friedman test found a significant difference among the three conditions, χ^2(2, N=14) = 23.21, *p* < .001. Follow-up pairwise Wilcoxon tests indicate users were significantly less likely to want to use multimedia than either spatial (*z* = 1.14, *p* = .007) or progressive detail (*z* = 1.54, *p* < .001), with no significant difference in reported desire to use progressive detail vs. spatial.

At the end of the session, participants ranked their overall preferences for the spatial, multimedia, progressive detail, and standard techniques, with a rank of 4 indicating the most preferred technique down to 1 indicating the least preferred. Participants displayed similar patterns in their preference rankings for alt text styles, as indicated by a Friedman test (χ^2(3, N=14) = 28.03, *p* < .001). The mean rank values for each technique were progressive detail (3.79), followed by spatial (2.86), then standard (2.00), and lastly multimedia (1.36). All participants indicated that progressive detail was either their favorite or second favorite technique, with 78.6% indicating it was their favorite. Follow-up pairwise Wilcoxon tests show that progressive detail was consistently ranked higher than either multimedia (*z* = -2.43, *p* < .001) or standard (*z* = 1.79, *p* = .002), and that spatial was consistently ranked higher than multimedia (*z* = -1.50, *p* = .01); other pairwise comparisons were not statistically significant.

RQ2: Image Understanding

After experiencing each image, participants rated their level of agreement with the statement "I am confident that I understand the image" on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree). We computed a single confidence score for each user for each of the three conditions (spatial, multimedia, and progressive detail) by taking the median of their confidence rating across the four scenarios (textbook, social media, news, shopping) that they experienced using each technique. We conducted a Friedman test to compare the median confidence scores across the three conditions, finding a statistically significant difference, χ^2(2, N=14) = 16.48, *p* < .001. Follow-up pairwise Wilcoxon tests reveal that participants were significantly more confident they understood images when using progressive detail (median rating 4.5) as compared to either multimedia (median rating 3.5, *z* = 1.43, *p* < .001) or spatial (median rating 4.0, *z* = 0.93, *p* = .04); the difference in ratings between the multimedia and spatial techniques was not statistically significant.

Additionally, after experiencing each image, participants used a three-point scale to rate the level of detail provided by the current alt text (-1 = too little detail, 0 = just enough detail, 1 = too much detail). We computed a single detail score for each user for each of the three conditions (spatial, multimedia, and progressive detail) by taking the median of their confidence rating across the four scenarios (textbook, social media, news, shopping) that they experienced using each technique. The median score for progressive detail was 0 (indicating the right amount of detail) whereas for spatial and multimedia it was –0.5 (indicating too little detail). We conducted a Friedman test to compare the median detail scores across the three conditions, finding a statistically significant difference, χ^2(2, N=14) = 7.32, *p* = .03. Follow-up pairwise Wilcoxon tests, however, do not indicate statistically significant pairwise differences (the difference between progressive detail and multimedia is only significant at *p* = .03 before Bonferroni corrections applied, but this drops to marginal significance (*p* = .09) after applying the corrections).

Table 2. Median scores (on a five point Likert-scale where 1 = strongly disagree and 5 = strongly agree) indicating participants' agreement with the statement, "I would use <technique X> when consuming images in <scenario Y>." Standard deviations are shown in parentheses.

	Digital Textbook	Social Media	News Website	Online Shopping
Progressive Detail	5.0 (0.5)	4.5 (1.1)	5.0 (0.8)	5.0 (0.4)
Spatial	5.0 (1.4)	5.0 (1.4)	5.0 (1.3)	5.0 (1.3)
Multimedia	2.5 (1.1)	3.0 (1.4)	4.0 (1.3)	2.0 (1.4)

RQ3: Pros and Cons of Novel Interactions

Feedback on Progressive Detail Alt Text

Participants' favorite aspect of the progressive detail prototype was the ability to choose how many and which levels of detail to listen to based on their interest in the image; nine participants (64.3%) mentioned this in their comments. P2 stated, "This is fantastic, I have flexibility. I can skip things if I want to." P5 commented liking how the technique "gave me the control of getting more information if I wanted to." P11 said, "[progressive detail] gave you a choice on how much information you wanted… I would use the feature 100% of the time." P13 contrasted the control offered by progressive detail to her everyday experience with standard alt text, observing, "I like to decide when I wanted more information. Today it's like all or nothing."

After experiencing all four scenarios using the progressive detail interface, we reminded participants that although our examples had three levels of detail, the format was flexible in terms of how many levels it could support; we then asked participants how many levels of detail they thought should be standard. The majority of participants (57.1%) felt three was the right number of levels to default to, although two wanted more (P8: "three to five, no more than five"; P11: "no less than three"). P4 liked the idea of just having a single, very detailed level of alt text, but also felt that offering three levels was okay. P6 felt there should not be a standard number of levels, noting that “it depends on the image itself.” P7, P9, and P15 preferred the idea of two levels, with the first brief and the second very detailed (and redundantly encapsulating the first, so that the user could simply choose if they wanted the shorter or longer version in a single interaction instead of progressively indicating that they want to listen to each level).

P2 pointed out that it may be important to create authoring guidelines about what types of detail to include at each level, noting that “details should be ordered carefully… how a sighted person sees an image [i.e., the order in which they notice details] should be the same way I hear it.”

Feedback on Spatial Alt Text

Ten participants (71.4%) commented that they enjoyed the experience of understanding the locations of objects within the image. P2 said, “[the spatial technique] showed me where people or items were in relation to each other… it made me appreciate what you guys generally see every day.” P5 commented, “I was able to get the detail about the picture itself and where everybody is in relation to the picture. It gave me a picture to visualize.” P6 said, “I like to know how objects are placed in [the] image to get an idea of where everything is. I’m able to build a mental map of the image.” P13’s favorite part was “seeing where everything was compared to everything else.”

The hands-on exploration style was very compelling for some users. P7 noted, “I like how you can explore the image [with fingers].” P9 said, “I do like moving my finger around.” P13 commented, “I am a very tactile person, so being able to get my hands on a picture was cool.” P7, however, noted that moving one’s hands over the image was not a time-efficient technique, pointing out that it could be “a lot of work to explore something.”

Four participants (28.6%) noted that the technique could be improved by having a greater density of objects labeled within each image (the images in our study had between two and five labeled regions). For example, in the social media scenario there was a photograph of a bride and groom, in which the bride and groom were each separately labeled; however, P3 wanted separate components of the people’s bodies labeled, for example “in things like [the] wedding picture, more things about what they were wearing.” P11 noted that it was desirable that the labels “include all small details.”

Sometimes, when using the spatial alt text, participants did not discover all of the labeled regions, and therefore did not hear all of the available descriptive information. For example, when exploring the image of Mount Rushmore for the textbook scenario, P2 touched only two of the four labeled presidents’ faces, although she correctly guessed that it was an image of Mount Rushmore from this partial information. Some participants expressed concern about the possibility that they might miss information during their exploration; for instance, P12 felt that one drawback of this technique was “not knowing how many things are tagged” within the image. Four participants (28.6%) suggested that it might be helpful to provide a brief caption giving an overview of the entire image in combination with the spatial exploration technique to help ensure that users would understand the key takeaways about an image even if they missed a labeled item during the direct exploration. P1 noted the importance of providing a context-setting overview caption by using an analogy to a puzzle: “[the spatial labels are like] a puzzle that need [sic] great imagination… needs something that connects each piece.” P7 noted that the context of supplementing the spatial labels with an overview caption would also help a user determine if they were sufficiently interested in an image to want to explore it spatially.

Feedback on Multimedia Alt Text

Of the three techniques tested, multimedia received the most varied feedback. Six participants (42.8%) felt that the non-verbal sounds did not add substantial value to their experience of consuming images. P1 described the alt audio as “silly,” and P11 felt that the extra sounds “may become more of a distraction.” P10 noted annoyance when he felt that “the sounds didn’t relate to the image text,” and P1 also felt that “the sounds doesn’t [sic] really relate to the images”; this feedback suggests that a well-crafted alt audio may be worth including, but that this property may be less applicable for general use for all images than some of the other interaction styles. For example, P13 suggested, “I like the idea of it in a situation when the sound comes directly from the event [depicted in a photo].”

In contrast, five participants (35.7%) noted that they particularly enjoyed the multimedia alt text because of its evocative nature. P1 mentioned that “the sound supported the meaning of the picture, which lead to an emotional reaction” and P7 commented, “[multimedia] makes browsing images more interactive, [it] draws an emotional attachment to an image.” P11 also noted the impact of the alt audio, noting that “the sound can be associated with emotion.” P12 commented that the multimedia effects were “funny” and made her smile. P4 appreciated “the potential for the background audio to give [a] kind of ambience.”

Four participants (28.6%) mentioned that the multimedia alt text interfered with their ability to comprehend the primary alt text due to the simultaneous presentation of music or sound effects with the verbal content. P4 said, “Having the text and the audio with text was hard to understand.” P12 noted, “the sound got in the way of hearing the verbal description.” P13 commented, “I don’t like that it [the music or sound effects track] interferes with the screen reader.” However, some participants felt the sound improved their comprehension of the image. P6 noted that the multimedia interaction helped him envision the images more richly, commenting that “the sound made the image livelier and brought more of a [sic] imagination to the picture.” Similarly, when experiencing the social media scenario image of college students celebrating at graduation, P10 said, "[the] cheering noise gave me more context of the image." Each time they experienced an image with the multimedia alt text, we asked participants how the alt audio component impacted their understanding of the image: in 69.6% of cases participants said there was no impact, in 12.5% of cases they said the audio decreased their comprehension, and in 17.9% of cases they felt it increased their comprehension.

DISCUSSION

This research is a first step toward enhancing the experience of digital imagery for screen reader users. Our taxonomy of caption properties, prototypes embodying new combinations of properties from that design space, and user feedback on those prototypes yields insights into potential benefits and challenges of supplementing the status quo experience of hearing a simple alt text. In this section, we discuss the implications of this work for creating rich representations of visual content for screen reader users.

Designing Alt Text

Creating a taxonomy of five key attributes of image captions (*interactivity*, *stability*, *representation*, *structure*, and *personalization*) helped us to conceptualize novel alt text styles (progressive detail, spatial, multimedia, categories, question-and-answer, and hyperlink). Of course, the six techniques included in our prototype are not the only possibilities for alt text redesign; our design space can serve as a blueprint to help ideate novel alt text possibilities by creating new contributions of these five attributes.

Our analytical method for devising the taxonomy, and our goal in proposing the taxonomy as a foundation upon which others might expand, is inspired by prior work such as "The Design Space of Input Devices," [8] in which Card et al. analytically identify key properties of input hardware and illustrate how new combinations of these properties might suggest novel input techniques. As in Card et al.'s work, our taxonomy was not derived from a design framework, per se, but rather based on a thoughtful analysis of the domain. Of course, it is impossible to prove that such taxonomies are complete – indeed, as interactive technologies evolve, we suspect the taxonomy of the design space for non-visually rendering image content will grow and evolve, as well. We found the current taxonomy to be a useful tool for designing and analyzing alt text – it is a starting point in a design conversation that we hope other researchers join.

Authoring Alt Text

Currently, missing or low-quality alt text is a pervasive problem [5, 13, 17, 18, 21, 22, 25, 29, 34], and it is unclear how our prototype supplements to alt text might impact compliance with alt text guidelines. One might hypothesize that the additional XML attributes that many of our alt text styles require could further reduce compliance by adding to content authors' burden. Conversely, one might hypothesize that some of our alt text styles might support improved compliance and quality. For instance, highly structured alt text (exemplified by our categories prototype) might benefit human authors or AI algorithms (structured templates have been shown to increase alt text quality for STEM diagrams [20]). Similarly, alt text that evolves rather than remain static (as exemplified by our question and answer prototype) allows end-users to improve missing or low-quality descriptions by interacting with content authors, bots, or crowds to give feedback on or request more information about captions.

An important area for future work is not only to measure the impact of different alt text formats in terms of authoring time, effort, and compliance, but also to develop authoring tools that can support low-effort production of high-quality alt text, such as WYSIWYG editors for labeling image regions for the spatial or hyperlink styles.

Consuming Alt Text

In our evaluation, we focused on subjective metrics (e.g., preference, confidence); however, before deciding on a new alt text standard, it will be necessary to perform assessments using objective metrics, as well (e.g., time spent, ability to answer questions about the image, ability to describe the layout of the image, ability to complete tasks in situ). Our focus on subjective metrics was appropriate for a first study of this space, and yielded initial insights useful for understanding attitudes and refining techniques. Further studies using different metrics, as well as evaluations in more realistic scenarios (e.g., longer-term use) will be important for increasing our understanding of this new area.

It is also important to bear in mind that improvements in accessibility for one audience could create new accessibility barriers for other groups. For example, the incorporation of non-speech audio in our multimedia prototype created a more immersive and emotive experience for some users, but this interaction might be unappreciable by someone who is deaf-blind and relies on a refreshable braille display to consume their screen reader's output. Accessibility issues related to economic class are also a concern; a technique that requires very expensive computing equipment (e.g., a special tablet with spatially-localized haptic feedback) may exclude participation by users who cannot afford such a device.

Also, our evaluations focused on only three of our prototype interactions; further work is necessary to understand the pros and cons of the *hyperlink*, *Q&A*, and *categories* interaction styles.

Beyond Digital Images

The need to label visual content extends beyond software and digital documents to the physical environment. While not alt text *per se*, creating digital descriptions of physical content is a growing area of interest. The VizWiz application uses crowd workers or social network contacts to caption smartphone photos of a user's surroundings [4, 6]. RegionSpeak [42] uses crowed workers to label semantically meaningful locations on an image of an inaccessible physical interface (e.g., annotating the locations of individual buttons and knobs on an image of an appliance). Eyes-Free Art [26] supplements previously inaccessible paintings in museums with proxemic interactions, in which a depth camera is used to measure the distance from a viewer to a painting and play different types of audio description depending on the viewer's distance (e.g., a verbal description of the painting, music that reflects the painting's genre, sound effects of objects in the painting). In this work, we showed how rich interactions can enhance the consumption of imagery in digital media, including interaction styles partially embodied in systems for exploring the physical environment like VizWiz, RegionSpeak, and Eyes-Free Art. While our focus was on creating descriptions for digital images, extending our taxonomy and interactions to create rich interactive labels for augmented reality or virtual reality is an avenue for future investigation.

CONCLUSION

In this paper, we argued that the status quo experience of alt text, a standard that is more than two decades old, does not take advantage of the capabilities of modern computing technologies that could be used to provide a rich, immersive, and evocative experience of digital imagery for people who are blind. We articulated a taxonomy comprising five categories (*interactivity*, *stability*, *representation*, *structure*, and *personalization*) that can be used to create richer representations of visual content for screen reader users. We then introduced prototypes demonstrating six new experiences that supplement or transform standard alt text by combining different properties of this new design space. Finally, we presented detailed feedback on three of these novel "alt text" interactions from fourteen screen reader users.

REFERENCES

1. Adams, D., Kurniawan, S., Herrera, C., Kang, V., and Friedman, N. Blind Photographers and VizSnap: A Long-Term Study. *Proceedings of ASSETS 2016*.
2. Berners-Lee, T. (1995). Image: IMG. In Hypertext Markup Language - 2.0 Document structure. World Wide Web Consortium. https://www.w3.org/MarkUp/html-spec/html-spec_5.html#SEC5.10
3. Berners-Lee, T. and Connolly, D. Hypertext Markup Language - 2.0. November 1995. Retrieved February 24, 2017 from https://tools.ietf.org/html/rfc1866
4. Bigham, J.P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R.C., Miller, R., Tatarowicz, A., White, B., White, S., and Yeh, T. VizWiz: Nearly Real-time Answers to Visual Questions. *Proceedings of UIST 2010*.
5. Bigham, J.P., Kaminsky, R.S., Ladner, R.E., Danielsson, O.M., and Hempton, G.I. WebInSight: Making Web Images Accessible. *Proceedings of ASSETS 2006*.
6. Brady, E.L., Zhong, Y., Morris, M.R., and Bigham, J.P. Investigating the Appropriateness of Social Network Question Asking as a Resource for Blind Users. *Proceedings of CSCW 2013*.
7. Brady, E., Morris, M.R., and Bigham, J.P. Gauging Receptiveness to Social Microvolunteering. *Proceedings of CHI 2015*.
8. Card, S., Mackinlay, J.D., and Robertson, G.G. The Design Space of Input Devices. *Proceedings of CHI 1990*.
9. Edwards, J. "Planet Selfie: We're Now Posting a Staggering 1.8 Billion Photos Every Day." *Business Insider*, May 28, 2014.
10. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., dollar, P., Gao, J., He, X., Mitchell, M., Platt, J.C., Zitnick, C.L., and Zweig, G. From captions to visual concepts and back. *Proceedings of CVPR 2015*.
11. Frolich, D. and Tallyn, E. Audiophotography: Practice and Prospects. *Extended Abstracts of CHI 1999*.
12. GSA Government-wide Section 508 Accessibility Program. http://www.section508.gov
13. Goodwin, M., Susar, D., Nietzio, A., Snaprud, M., and Jensen, C.S. 2011. Global web accessibility analysis of national government portals and ministry web sites. *Journal of Information Technology and Politics*, 8(1), 41-67.
14. Harada, S., Sato, D., Adams, D.W., Kurniawan, S., Takagi, H., and Asakawa, C. Accessible Photo Album: Enhancing the Photo Sharing Experience for People with Visual Impairment. *Proceedings of CHI 2013*.
15. Issa, Y.B., Mojahid, M., Oriola, B., and Vigouroux, N. Accessibility for the blind: An automated audio/tactile description of pictures in digital documents. *Proceedings of ACTEA 2009*.
16. Korpela, J. (September 1998). Guidelines on ALT texts in IMG elements. Retrieved 26 August, 2017 from http://www.cs.tut.fi/~jkorpela/html/alt.html.
17. LaBarre, S.C. 2007. ABA Resolution and Report on Website Accessibility. *Mental and Physical Disability Law Reporter.* 31(4), 504-507.
18. Loiacono, E.T., Romano, N.C., and McCoy, S. 2009. The state of corporate website accessibility. *Communications of the ACM,* 52(9), September 2009, 128-132.
19. MacLeod, H., Bennett, C.L., Morris, M.R., and Cutrell, E. Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images. *Proceedings of CHI 2017*.
20. Morash, V.S., Siu, Y-T., Miele, J.A., Hasty, L., and Landau, S. Guiding Novice Web Workers in Making Image Descriptions Using Templates. *ACM Transactions on Accessible Computing*, 7(4), November 2015.

21. Morris, M.R., Zolyomi, A., Yao, C., Bahram, S., Bigham, J.P., and Kane, S.K. “With most of it being pictures now, I rarely use it”: Understanding Twitter’s Evolving Accessibility to Blind Users. *Proceedings of CHI 2016*.
22. Olalere, A. and Lazar, J. 2011. Accessibility of U.S. Federal Government Home Pages: Section 508 Compliance and Site Accessibility Statements. *Government Information Quarterly*, 28(3), 303-309.
23. Petrie, H., Harrison, C., and Dev, S. Describing Images on the Web: A Survey of current Practice and Prospects for the Future. *Proceedings of HCI International 2005*.
24. Pilgrim, M. The longdesc lottery. September 14, 2007. https://blog.whatwg.org/the-longdesc-lottery
25. Power, C., Freire, A., Petrie, H., and Swallow, D. Guidelines are only half of the story: Accessibility problems encountered by blind users on the web. *Proceedings of CHI 2012*.
26. Rector, K., Salmon, K., Thornton, D., Joshi, N., and Morris, M.R. (2017) Eyes-Free Art: Exploring Proxemic Audio Interfaces for Blind and Low Vision Art Engagement. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technology (IMWUT 2017)*.
27. Salisbury, E., Kamar, E., and Morris, M.R. Toward Scalable Social Alt Text: Conversational Crowdsourcing as a Tool for Refining Vision-to-Language Technology for the Blind. *Proceedings of HCOMP 2017*.
28. Slatin, J.M. and Rush, S. Maximum Accessibility: Making your Web Site More Usable for Everyone. *Addison-Wesley Longman Publishing Co., Inc.* 2002.
29. Shi, Y. E-Government Web Site Accessibility in Australia and China: A Longitudinal Study. *Social Science Computer Review*, Fall 2006. 24: 378-385.
30. Stolley, K. Image Accessibility Part I: Beyond alt Attributes. *Digital Rhetoric Collaborative*. June 2016. http://www.digitalrhetoriccollaborative.org/2016/06/15/image-accessibility-part-i-beyond-alt-attributes/
31. Tran, K., He, X., Zhang, L., Sun, J., Carapcea, C., Thrasher, C., Buehler, C., and Sienkiewicz, C. Rich Image Captioning in the Wild. *Proceedings of CVPR 2016*.
32. Twitter Blog. Accessible Images for Everyone. March 29, 2016. https://blog.twitter.com/official/en_us/a/2016/accessible-images-for-everyone.html
33. von Ahn, L., Ginosar, S., Kedia, M., Liu, R., and Blum, M. Improving accessibility of the web with a computer game. *Proceedings of CHI 2006*.
34. Voykinska, V., Azenkot, S., Wu, S., and Leshed, G. How Blind People Interact with Visual Content on Social Networking Services. *Proceedings of CSCW 2016*.
35. Web Accessibility in Mind (WebAIM). Alternative Text. http://webaim.org/techniques/alttext/
36. World Health Organization. “Visual Impairment and Blindness” factsheet, August 2014. http://www.who.int/mediacentre/factsheets/fs282/en/
37. World Wide Web Consortium (W3C). Accessible Rich Internet Applications (WAI-ARIA) 1.1. https://www.w3.org/TR/wai-aria/
38. World Wide Web Consortium (W3C). HTML5 Image Description Extension (longdesc). https://www.w3.org/TR/html-longdesc/
39. World Wide Web Consortium (W3C). Techniques for WCAG 2.0, H37: Using alt attributes on img elements. http://www.w3.org/TR/WCAG20-TECHS/H37.html
40. World Wide Web Consortium (W3C). Web Accessibility Tutorials: Images Concepts. https://www.w3.org/WAI/tutorials/images/
41. Wu, S., Wieland, J., Farivar, O., and Schiller, J. Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service. *Proceedings of CSCW 2017*.
42. Zhong, Y., Lasecki, W.S., Brady, E., and Bigham, J.P. RegionSpeak: Quick Comprehensive Spatial Descriptions of Complex Images for Blind Users. *Proceedings of CHI 2015*.